DOCUMENT RESUME

ED 294 215 CS 211 200

AUTHOR Haswell, Richard H.
TITLE Contrasting Ways To Appraise Improvement in a Writing Course: Paired Comparison and Holistic.
PUB DATE Mar 88
NOTE 22p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (39th, St. Louis, MO, March 17-19, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College English; *Evaluation Methods; Freshman Composition; Higher Education; *Holistic Evaluation; Scoring; Testing Problems; Writing (Composition); *Writing Evaluation; Writing Research
IDENTIFIERS Evaluation Research; *Intrasubject Paired Comparison

ABSTRACT

To compare the different images of writing that different assessment methods produce, a study examined two formal writing assessments--holistic and the specially developed intra-subject paired comparison method (IPC)--of pre/post university freshman composition-course writing. The samples of writing were unrehearsed, 50 minute, in-class essays. Forty students were randomly selected from freshman composition courses to write on pre/post switched topics. Essays were evaluated by both IPC and holistic methods. IPC differs from holistic evaluation by making an analytical evaluation of two essays written by the same student (one early and one late in the course), comparing separate writing subskills (ideas, support, organization, diction, syntax, mechanics, and overall quality). Findings indicated that holistic evaluation was more costly (44 rater-hours as opposed to 33 for the IPC) and time-consuming. Both methods seemed equally sensitive in detecting overall improvement. However, the IPC method recorded more individual improvement because it rated subskills separately. (Six figures and one table of data are included.) (MM)

Richard H. Haswell
Department of English
Washington State University
Pullman, WA 99164—5020

Contrasting Ways to Appraise Improvement in a Writing Course: Paired Comparison and Holistic

"The altering eye alters all"
—William Blake

Blake's old truth, that the way we go about seeing something alters what we see of it, is one we know full well but forget full often. It is a truth that those involved in evaluation of writing must think about constantly. As new methods of assessment slowly and surely proliferate, we ought to be comparing the different images of writing that the different methods produce.

It is one such comparison that I will present here. I will describe two different formal assessments of one sample of pre/post university-freshman, composition-course writing, and will compare the results. Of pragmatic interest is how the two methods compare in terms of cost, reliability, and validity. Of theoretical interest is how the two distinct methods of evaluation generate different pictures of the same pieces of writing. Each picture records some outcomes which the other misses.

The general aim of this report is to introduce a new procedure of assessing the effect of a writing course on students—a procedure I will call Intrapersonal Paired Comparison. The report, both the study and its findings,

should be considered preliminary. Much of it was presented at the annual College Composition and Communication Convention in St. Louis, April, 1988.

Background

Since one of the two assessment procedures—the holistic—is familiar, I will start by describing the other. I should begin by noting that the circumstances for developing this new procedure grew, nonetheless, out of a familiar situation, namely English department paranoia—that periodic crisis formed of mutterings from some department members that the freshman writing course is a waste of resources, of committee-room complaints from the agronomy faculty that English is not teaching their students to write, of informal concerns expressed formally by deans, etc. In sum, this new procedure grew out of a political need to show that students in an English department's one-semester freshman writing course (at a large, land-grant state university) really did improve their writing.

At the time I was aware that holistic methods of assessing writing improvement sometimes had not detected that improvement in similar courses in the past, or had placed improvement not very convincingly within the parameters of acceptable statistical confidence. I feared, as others had, that holistic scales were not fine enough to capture the small but meaningful progress which many teachers see their students achieving during the course. I also knew that a major problem with holistic information is that it provides, finally, little that is diagnostically useful to teachers. Compared to first semester essays, end-of-the-semester may rise significantly 1.2 points on a 6-point scale, but that tells us nothing about what students have learned or not

learned (Edward M. White summarizes these problems well in Chapter 2 of Teaching and Assessing Writing, San Francisco: Jossey-Bass, 1986).

So I devised the following scheme. It hopes to measure even small increases in individual improvement by forcing a comparison between two essays, one composed early and the other late in the course. It departs from the holistic method by making this comparison analytically, between separate writing subskills. It also departs from the holistic in that it directly compares the writings of the same student, not of different students. For that reason I will call the method "intra-subject" paired comparison, or IPC for short.

The Intra-Subject Paired Comparison Method (IPC)

The IPC evaluator rates essays in batches of two. The rater knows that each pair of essays was written by the same student, pre and post compositions written on switched topics, but does not know which is pre and which post. On the scoring sheet for each pair of essays (Figure 1, p. 4), then, will be recorded the rater's impressions not of one essay but of two essays. The rater's first task is to compare the companion essays in terms of Ideas or content. If the left-hand essay (position is awarded by chance) is greatly better in terms of its ideas, the left box marked "GB" is marked; if just obviously better, the left box "OB"; if only a little better, even if only a minim better, the left box "LB." If, on the other hand, the right-hand essay is a touch better, or obviously so, or greatly so, then the appropriate boxes on the right will be marked. Note that the two essays are ranked, then, in terms of ideas. Also note that the "Little Better" box fits the situation where a rater can only intuit a difference. There is no box letting the rater off the hook by declaring that no difference exists between the two essays in any aspect. Such

Figure 1

INTRA-SUBJECT PAIRED COMPARISONS

	Essay A _____			Essay B _____		
	Greatly Better	Obviously Better	A Little Better	A Little Better	Obviously Better	Greatly Better
Ideas						
Support						
Organization						
Diction						
Sentences						
Mechanics						
Overall						
Passing Level	Yes ☐			Yes ☐		

forced-choice comparisons have been used before—for instance in Andrew Kerek, Donald A. Daiker, and Max Morenberg's 1976-7 University of Miami study (see Sentence Combining and College Composition, Monograph Supplement 1-V51 of Perceptual and Motor Skills, 1980, pp. 1109, 1117-1119)—but not, as far as I know, between subskills of pieces of writing produced by the same writer.

The rater continues the assessment by making a similar comparison in terms of Support, then Organization, Diction (or word choice), Sentence Structure (or syntax), and Mechanics (or surface error). This categorization of subskills, incidentally. is based in part on Paul B. Diederich's factoring of

teacher responses to writing (there is a convenient summary in Chapter 2 of Measuring Growth in English, Urbana: National Council of Teachers of English, 1974). My first four subskills repeat his factors. But in place of what he calls "flavor," I have Sentence Structure and Mechanics, as categories more easily distinguished by raters and more useful for teachers.

The rater ends by making two more comparisons. One is in terms of overall quality of the essay. This is an acknowledgment of the holistic premise that the artistic whole of an essay may be greater than the sum of its writerly parts, or different than an averaging of these parts. Finally, the rater judges, separately for each essay, whether the essay is of passing quality for the course. This is an acknowledgment that the IPC scheme fails, rather blatantly, to be criterion referenced. It judges whether a sample of one student's writing has improved or regressed from an earlier sample, and roughly how much it has improved or regressed. But it does not judge the quality of either sample relative to any outside standard. Take one rater's assessment of one pair of essays (Figure 2, p. 6). The writing here may represent a student progressing from F to D work, or from B to A work.

In our original evaluation using the IPC method, each pair of essays was assessed this way by three trained raters, working independently. Figure 3 (p. 6) shows a scoring sheet combining these three assessments and determining the final assessment for this pair. (The samples of writing, it should be understood, were unrehearsed, 50-minute, in-class essays.) The plain X's represent the independent judgments, the circled X the final decision. Notice that here, with two raters selecting Essay B as a little better organized than Essay A, but another rater selecting Essay A as obviously better, a fourth rater was needed. Essays were re-read when there was one Obviously

Figure 2

	Essay A			Essay B		
	Greatly Better	Obviously Better	A Little Better	A Little Better	Obviously Better	Greatly Better
Ideas				X		
Support					X	
Organization				X		
Diction				X		
Sentences				X		
Mechanics					X	
Overall				X		
Passing Level	Yes			Yes	X	

Figure 3

	Essay A PRE			Essay B POST		
	Greatly Better	Obviously Better	A Little Better	A Little Better	Obviously Better	Greatly Better
Ideas			X	XX⊗		
Support			X	⊗X	X	
Organization		X		XX⊗	X4	
Diction				XXX⊗		
Sentences				X	XX⊗	
Mechanics			X	XX⊗		
Overall			X	XX⊗		
Passing Level	Yes	X		Yes	XXX	

Better or Greatly Better mark on one side with other marks on the other side. Essay B, in fact, is the post essay: a fact the teacher in us is happy to see.

It may help to visualize the IPC procedure by comparing its assessment of the two essays rated in Figure 3 with the assessment recorded by an independent holistic procedure (a general comparison of these two assessments will be taken up below). When these two essays were read holistically, independently of each other, they achieved the identical summed score, a score in the lowest category of the holistic scale. So the holistic shows what this part of the IPC could not, that these are two very poor essays; and the IPC shows what the holistic could have shown but did not, that in some ways the end-of-the-semester essay is an improvement over the beginning essay. That the holistic rating finds it hard to distinguish differences among very poor essays is a point I will return to. For the moment, notice another difference here. Since the lowest category of the holistic assessment was defined as "failing," the raters looking at these two essays through the holistic method must have seen pieces of writing that would not pass the course. But the raters looking at them through the IPC method saw only one as failing, the other as passing. (The student, incidentally, got a C+ for the semester.)

Only 12 of the 40 students we assessed showed, as in Figure 3, course improvement in all 6 subskills. Another pair of essays is more typical (Figure 4, p. 8). Essay B is the end-of-the-semester piece. The holistic rated it a complete failure, all 4 raters giving it the bottom rate (summed score = 4). They rated Essay A here as weak but passing, giving it a score 2 1/2 times as high as Essay B (summed score = 10). The IPC evaluation, on the other hand, saw Essay B as better in the majority of subskills and, perhaps consequently, as the better essay overall.

Figure 4

	Essay A PRE			Essay B POST		
	Greatly Better	Obviously Better	A Little Better	A Little Better	Obviously Better	Greatly Better
Ideas			X	XX⊗		
Support		X	XX⊗			
Organization		XX⊗		X		
Diction			X	XX⊗		
Sentences				XXX⊗		
Mechanics				X	XX⊗	
Overall			X	X⊗	X	
Passing Level		Yes X			Yes XX	

Results of the IPC Assessment

Before proceeding to a full comparison of the holistic and the IPC methods, it is worth showing how the IPC assessed the composition course as a whole. We randomly selected 40 students from a number of sections of the course to write on pre/post switched topics. More precisely, all students in the freshman course wrote on the pre topic during the second meeting of class, but only four sections were selected, toward the end of the semester, to write on the post topic. The choice of sections was random, but it was checked to make sure we had a reasonable representation in terms of teacher experience and skill. Nineteen percent of the students in these sections failed to take the post-test. It can be argued that this attrition may have helped the

Figure 5

EVALUATION OF FRESHMAN COMPOSITION

PRE	First-of-the-semester essays (N = 40)
POST	End-of-the-semester essays (N = 40)
GB	Greatly better
OB	Obviously better
LB	A little better

	PRE			POST			Comparisons favoring POST essays	Sign test
	GB	OB	LB	LB	OB	GB		
IDEAS		5 %	30%	48%	18%		66%	< .05
SUPPORT		13%	38%	33%	18%		51%	NS
ORGANIZATION		15%	25%	35%	25%		60%	NS
DICTION		5%	23%	63%	8%	3%	75%	< .01
SENTENCES			33%	60%	8%		68%	< .05
MECHANICS			25%	55%	18%	3%	76%	< .01
COMPOSITE		6%	29%	49%	15%	1%	65%	< .001

PASSING LEVEL

YES	XXX	52%	XXX	85%
YES	XX	37%	XX	10%
YES	X	8%	X	5%
YES		3%		0%

ultimate finding of writing improvement in the course, but it should be remembered that generally the poorest students progress the most during a writing course.

Figure 5 (p. 9) presents the IPC summary data. Pre essays are on the left, post on the right. The percentages here show how often a category was picked among the 40 pairs of essays, so rows add up to 100%. To test statistically for overall improvement, at least two procedures are appropriate and familiar. Individual preferences for pre or post in all categories might be summed, and a sign test administered: here IPC found 156 out of 240 choices (65%) favoring post essays ($p < .001$). Or a chi-square could be run, which here produced a χ^2 of 26.61 ($p < .05$). For individual subskills, the less powerful statistic was applied here, the sign test, which simply uses the number of times a post essay was judged an improvement at all (collapsing the LB, OB, and GB distinctions) over a pre essay, in each category. The sign test has the advantage of being easily gasped and easily computed (one hardly needs even a hand calculator). As Figure 5 indicates, 4 of the 6 categories show significant improvement.

For teachers of the course, the diagnostic information here was enlightening, to say the least. In two areas where they expected to get improvement—in organization and in support—they did not; but in areas where they thought they had little effect, they did—in ideas, vocabulary, and syntax. Added support for the course may be seen in the "Passing Level" decisions. Nearly half of the pre essays, in contrast to only 15% of the post essays, earned one or more rates of fail from the three raters. There is one other result that does not show up on the summary sheet in Figure 5: the "Overall" decision in effect was superfluous. Only once with the 40 pairs of essays did the "overall" judgment reverse the trend of the categories. It may be pointed

out too that the "Greatly Better" category did not add much information. It was picked only 38 times out of 737 individual rater choices (5%), and ended as a final decision only 2 times out of 280. It was, however, a category that evaluators said they appreciated having available.

Results of a Holistic Assessment of the Same Essays

Later, these 80 essays were rated by a formal holistic assessment. Raters, none of whom had participated in the IPC assessment, were trained on a holistic ranking with 4 basic categories (low or failing, medium low, medium high, and high), each category divided to make an 8-point scale (Figure 6). Essays of course were not read in pairs but individually. Each

Figure 6

HOLISTIC RATING SCALE

HIGH	**8** **7**
MEDIUM HIGH	**6** **5**
MEDIUM LOW	**4** **3**
LOW (FAILING)	**2** **1**

essay received 4 independent ratings, generating a possible range of scores from 4 to 32. Inter-rater reliability was .88 (Cronbach's alpha); the average correlation between two raters was .64.

Now if we use these holistic scores to assess improvement in the course, we in fact get it. Pre essays averaged 16.2 on the scale, post essays 19.4. A correlated t-test finds this difference significant at the .01 level of confidence (p .005, DF 39, t -3.01). I think it should be pointed out that this happy result is achieved in part by using 4 independent raters and an 8-point scale, which produced a wide spread of data points. If we reconvert the data to the more customary 3-rater and 4-point scale system, the difference between pre and post essays is much less convincing (7.0 and 7.9), a difference that barely scrapes by under the .05 level of confidence (p .039, DF 39, t -2.14). Here the degree of improvement recorded during a one-semester freshman course and the results from statistical confidence-testing are comparable to similar holistic assessments in the past.

Comparison of the IPC and the Holistic Assessments

We are now in a position to compare these two systems of assessment. In terms of cost, the holistic was more expensive, 44 person-hours as opposed to 33 for the IPC. The holistic used 8 readers requiring 3 hours of training and 2.5 hours to rate the 80 essays (I am not calculating the time spent developing the anchor essays). The IPC with 6 readers took 1.5 hours of training and 4 hours rating. Readers averaged 2 minutes assessing each essay holistically (about average, according to White), and 4 minutes for each essay by forced comparison. The time saved for the IPC was in the training, where suitable

concordance was achieved in comparing a subskill of the two essays much more quickly than in placing an essay in a holistic category.

Perhaps, however, this training should have been more extensive. It looks as thought rater reliability was lower for the IPC. If we treat the six IPC choices as ranks (Pre "GB" as lowest, Post "GB" as highest), just as in a holistic ranking, and then run correlations between pairs of raters, the median correlation hovers around .50 (compared to .65 for the holistic). One reason this is so low is because very rarely were the two extreme ranks (GB) chosen, and it is difficult to get a high correlation on a scale of only four. On the other hand, with 3 raters, a choice had to be submitted to a fourth reader only 9.6% of the time, which would be, according to White, a reasonable reliability on the holistic. This meant that 15 out of the 40 pairs of essays had to be re-read (often to decide on only one or two subskills, of course). In terms of categories, something of the relative difficulty in getting raters to agree can be seen by looking at how many instances of each subskill required a re-reading: Support 9 times, Organization 6, Ideas and Sentences 5 each, Mechanics 2, Diction and Overall none. I would hazard that the IPC rater reliability can be raised considerably with a better organized training session, in particular with distinctions among the ranks (LB, OB, GB) more precisely made.

Incidentally, it can be argued that handwriting must influence the IPC decisions much less than holistic decisions, at least to the extent that the holistic is norm and criterion referenced. Since the IPC measures value distance between two writing samples from the same student, handwriting effects should balance out. But with the holistic, handwriting will influence where a particular essay stands in relation to other essays in the sample. This perhaps will balance out in calculating individual student progress by

comparing the pre and post holistic scores, but may affect how these essays stand in relation to any outside criterion (as where, for example, a score of 2 reflects passing level).

The two methods seem equally sensitive in detecting overall improvement in the course. Both methods found exactly the same percent of students advancing from pre to post: 68%, or 27 out of 40 (comparable to other assessments of freshman composition). Individually, however, the IPC method recorded more improvement, in part because it rated sub kills one by one instead of all together. So whereas the holistic found 9 students regressing from pre to post (4 others earned the same post holistic summed score as pre), the IPC recorded only 3 students regressing in all six categories, the other 37 students showing improvement in some aspect of their writing. The IPC also may be more sensitive to improvement at the ends of the quality spectrum, with the very poor and very good writers. We have already seen several examples of poor essays showing little difference holistically but a substantial difference by forced comparison. The same seems to be true at the upper end. Table 1 (p. 15) compares the two methods, dividing the 40 students into quartiles by initial writing ability (as judged on the holistic). To compute course progress in Table 1, for the holistic the sum of holistic rates was used, and for the IPC the summed accomplishment on all six of the subskills, with a count of 1 awarded for a decision of LB, 2 for OB, and 3 for GB. The holistic pattern—where the worse the writer stands initially, the more improvment that writer records—is a common finding in such evaluations. But the IPC shows the top quartile of students gaining as much as the bottom quartile, and the medium high or "B" student most unlikely to record gain. The two methods obviously are discovering improvement—or more

basically, visualizing quality in writing—in some importantly different ways.

Some of these differences become obvious with a look at individual cases. For one difference, the holistic raters seem to have been more influ-

Table 1

Initial holistic summed Score (4 raters)	Mean pre/post difference on holistic	Mean pre/post difference on IPC
7-12 (N = 11)	+5.82	+2.27
13-16 (N = 9)	+5.56	+4.56
17-20 (N = 11)	+3.27	-0.18
21-28 (N = 9)	-3.11	+2.11

enced by mere number of words in an essay. There were 12 pairs of essays where one essay is conspicuously shorter—over half a hand-written page shorter—than its companion. The holistic gave the longer essay a better rate in every case, by an average of 6.6 points (which is considerable, considering there were only 24 points on the scale). The IPC also preferred the longer essay, in 47 out of 72 subskill category choices, but discovered then some better writing qualities in the shorter piece 25 out of the 72 times. For four pairs out of the 12, the IPC awarded preference to the shorter essay in the majority of the categories and for one pair in an equal numer of categories. Since 8 of the 12 shorter essays were written at the beginning of the semester,

this is an situation where the holistic may be more likely to record course progress than will the IPC, but it is a likelihood achieved, perhaps, at the expense of validity.

For another difference, the holistic method seems to put more weight on the subskills of Support and Mechanics than does the IPC. In those essay pairs where the holistic records substantial quality gain and the IFC little gain, it is most frequently those two subskills which the IPC finds gain in. Even more common are the essay pairs where the holistic recorded little difference but the IPC substantial improvement. In seven instances, the IPC showed gain in all subskills except Support. Generally it seems that without a strength in Support, holistic raters have trouble seeing other strengths (cf. Sarah W. Freedman, "How Characteristics of Student Essays Influence Teachers' Evaluations," Journal of Educational Psychology, 71, June 1979, 328-338).

The holistic failed 7 essays and the IPC 6, so the two systems seem equivalent in lenience. But only two of these failing essays were the same—the systems disagreed on 11. Here, a couple of patterns are clear. The IPC failed an essay which the holistic passed when the essay did not surpass its companion on any or on only one of the subskills. And the holistic failed essays that the IPC passed when the essay was comparatively weak in only one or two subskills, usually Support or Ideas. In applying that outside criterion of "passing quality," the IPC method is obviously affected by the conscious and direct comparison with the companion essay, swayed by the number of subskills showing comparative success or failure. The holistic, on the other hand—limited to a comparison, perhaps an intuitive comparison, of internal traits within one essay—seems especially swayed by the opposite situation, the powerful halo effect of one or two subskills.

One final comparison, I think, affords a special insight into the two methods. There were15 students where the two assessment techniques disagree on the presence of improvement or regression during the semester. Twelve of those 15 cases involved instances where the independent holistic rates showed the greatest variance, that is, where the four holistic raters had the greatest trouble agreeing among themselves. Typically involved are essays that the IPC shows radically divided in subskill strengths, with a few skills showing strong improvement and a few showing strong regression. The implication is that the uneven essay, which students produce especially toward the beginning of the course, is more difficult for the holistic scheme to handle and easier for the IPC.

Procedural Strategies of Raters

It is instructive to consider these differences in light of the distinct ways these two assessment have raters proceed. The IPC raters must take up, compare, and rank subskills one by one until the six are completed. Holistic raters supposedly involve in their judgment all major factors of writing, but since this is not done systematically they may be more susceptible to the halo effect of one or more strong factors. This is not to say that the halo effect cannot occur in the IPC, and indeed the set order of taking up the six categories may well have produced a systematic halo effect of the first categories. One reason I consider the findings in the present study preliminary is that I see a need to test this particular IPC method by having different raters for each category, which should reduce halo effects.

A second essential difference in rater procedure lies in the constant comparison with another essay written by the same writer. The peculiar

influence of this method is totally unknown. Certainly the strain is less on the rater, partly because comparison is made with actual essays before the eye—essays carrying more features more akin and endowed with a directly meaningful relationship (two products of the same person). The idealism of the holistic method is hard on raters, especially in trying to reduce an uneven essay to an abstract holistic ranking with its neat hierarchy of categories. Operating here is the fact that the holistic usually is partly norm referenced and partly criterion referenced. The holistic rater then operates by setting up an individual piece of writing against at least two abstractions, the hierarchy of writerly values in the pre-set holistic scoring guide, cross-referenced with "course standards" at the lower categories of the scale. One reason why holistic raters are required to make decisions rapidly is because this procedure becomes more problematic the longer it is indulged in. The IPC, on the other hand, is neither norm or criterion referenced (that is, if we disregard the "Overall" and "Passing Level" decisions, which I consider non-essential to the procedure). If anything, it is self referenced. One piece of writing is set up against another.

A third difference has to do with the fact that the holistic rater must function, as the name says, holistically, while the IPC rater basically functions analytically. In the subskill decisions, the IPC rater has no need to weigh factors, factors sometimes quite removed one from another, to arrive at a summative judgment, but rather takes up each factor one at a time. Herein lies an advantage of the holistic, which may allow exceptional or original papers to work because, despite the exception or the originality, the whole works. One quality—say, a heavy focus on particulars—is allowed full rein and the assessment as a whole does not pay for it. So the holistic allows in the halo effect along with the individual or eccentric performance.

It is interesting to note that primary trait or performative assessment, which adds diagnostic information by concentrating on isolated traits, may punish exceptional works. Criteria are defined so precisely that an eccentric piece will get marked down (say, a piece with a one-sentence introduction). The halo effect is decreased and the diagnosis is improved, but individuality perhaps suffers. The IPC method lies somewhere between. It adds some diagnostic information, gives partial reward to exceptional effects, and curtails the halo effect.

Summary: IPC and Holistic

The methodology of the IPC recommends it for particular evaluative uses, which in turn suggest how the present preliminary form of it might be adapted and developed. Just as the holistic, with its testing of individual performances against an absolute system of writing values, seems best fit to compare groups (e.g., 14-year-olds against 17-year-olds) or to rank an individual within a group (e.g., placing an entering freshman), so the IPC, with its comparison of two performances of the same writer, seems best fit to assess the achievement of an individual within a course of instruction. This assumes—and it is not an assumption all teachers and administrators necessarily hold—that the essential function of a writing course is to foster improvement in writing. The IPC may stand as a method of assessment most amenable to writing teachers who are concerned less about the level of skill a student has on entering their course and the level or grade that student has earned on leaving it, and more about how much the student has progressed during it. Not only may administrators assess whole sections of a particular course by the IPC method (as was attempted in this study), but individual

teachers can assess their particular section. I have overseen some individual section-testing by means of the IPC, with beginning teaching assistants, and the results ranged from around 90% of the students achieving pre/post improvement in impromptu writing down almost to chance (50%).

Teachers, of course, can also compare beginning and end of the semester performance of their section through a holistic assessment, but certain questions will still remain unanswered: was the holistic sensitive enough to record a semester's progress, was the students' entering level of accomplishment higher or lower than normal, and diagnostically where lie the qualities of writing the course seems to have affected? In particular the holistic will not answer a question that seems both germane and pressing where teaching of novice writers is involved, namely whether advance in writing may have been non-lateral, whether indeed some aspects of writing may have not only progressed more rapidly in relation to others but also in despite or to the detriment of others.

The sensitivity and diagnostic specificity of the IPC suggest that it could not only helpfully assess course instruction as a whole, but also test particular components of the course, even particular lessons. One remembers that the essential method of the IPC, direct comparison of pre and post writing, has been used to evaluate instructional intervention under research conditions, for these same reasons of sensitivity and diagnosis. All this suggests one similarity of the IPC to the holistic, that both seem quite open to adaptation to particular circumstances. Just as the holistic can be modified along the lines of performative or primary-trait motives, so can the IPC. The number of comparative ranks could be modified (I suggest reducing the three here to two: an obvious difference in quality, and an intuited difference—but other situations might allow for an even more refined system than three). The

selection of subskills could be altered to match intentions of teachers or researchers or assignments, especially to answer questions about the interactions among instruction and various writing skills (e.g., does improvement in organization help or hinder improvement in support). And above all the pre/post writing tested could be rehearsed and drafted, better allowing both teachers and students a chance to succeed in the goals toward which the course may have worked most.

I hope this discussion has made it clear I am not arguing that either of the two evaluation systems is absolutely better. Compared to the holistic, the IPC does seem as capable and perhaps more easily capable of generating a convincing argument, certainly a more concrete argument, that a writing course fosters writing improvement. The IPC can do this with small number of essays and be as cost effective as the holistic. It can also do it without the time-consuming and highly technical task of constructing scoring guides and exemplary essays requisite for holistic, primary-trait, and performative assessments. But the holistic has its own virtues, and I am not recommending that the IPC or any other scheme replace it. My comparison here tends to support White's argument that there are vital differences in what an analytic approach, such as the IPC, and a holistic approach will see. Both seem to produce a viable appraisal of the essay as a whole, as the holistic doctrine urges, yet the vision of the two wholes seem to differ significantly as the eye alters from one method to the other Both methods, I think, deserve use because they are less systems of rating and more systems of appraisal, in that word's root sense of "finding praise or price." Both the holistic and the IPC, that is, help us see virtues, albeit different virtues, in student writing where we are otherwise apt to see defects, and I am for any method of evaluation which does that.